

Development of an AI-Driven Medical Diagnostic System for Type-2 Diabetes Detection

IGWILO CHIOMA GOODNESS¹, PROF V. E. EJIOFOR², DR. S. A. ALADE³, DR. ROSE U. PAUL⁴

^{1, 2, 3, 4}Nnamdi Azikiwe University (Unizik)

Abstract - Diabetes is a growing health concern worldwide, and early detection is crucial for effective treatment. This paper introduces a predictive model that combines Support Vector Machine (SVM) and Random Forest algorithms with a soft voting classifier to detect diabetes based on clinical data. The ensemble model leverages the strengths of both SVM and RF to improve overall performance. The methodology follows an Object Oriented Analysis and Design Methodology (OOADM) to model the system architecture and to address class imbalance issues, Synthetic Minority Over-sampling Technique (SMOTE) is employed as a data balancing technique. The proposed approach employs Electronic Health Records (EHR) and a user-friendly interface for efficient data analysis and real-time prediction incorporating attributes such as blood pressure, age, BMI, smoking status, and physical activity to predict diabetes. The results demonstrate the effectiveness of ensemble machine learning in identifying diabetes patients, achieving 96 % of accuracy, 97% of precision, 99% of recall, and 98% of F1-score highlighting its potential for early detection and providing a promising direction for future research and development of robust real-time prediction systems.

Keywords: Diabetes, machine learning, SVM, RF, Electronic Health Records(EHR)

I. INTRODUCTION

Diagnosis is described as both a process and a classification system used by the medical profession to identify and label a specific health condition. The detection of a patient disease, condition, or injury from its signs and symptoms is a process termed medical diagnosis. The accuracy and timely methodology of diagnosis offers the best opportunity in ensuring a patient positive health. Clinicians do not need to obtain diagnostic certainty prior to initiating treatment; the goal of information gathering which includes clinical history and interview, performing a physical examination, obtaining diagnostic testing, and consultation referrals in the diagnostic process is to reduce diagnostic uncertainty enough to make optimal decisions for subsequent care (Kassirer,2006).

Diabetes is a coarse disease in our society (Ajay *et al*, 2022) and is one of the most common diseases worldwide and its prevalence rate continues to rise. This increase is due to factors related to nutrition, lifestyle and genetic factors on the other hand, thus creating a real public-health problem. It is crucial to identify diabetes early in order to allow rapid treatment, capable of slowing down the progression of the disease (Nawal *et al*, 2023). Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis (Mujumbara *et al*, 2019). Nowadays youth is getting highly affected by Type-2 Diabetes Mellitus(T2DM) as reported in (Naz *et al*, 2020). Since diabetes has a massive effect on global health and the economy, it is essential to enhance methods for predicting and preventing diabetes (Chen's, 2022). The results of laboratory tests have an important impact on patients' care, as they influence physicians' decisions including admission, drug orders, and discharge as well as monitoring and managing the vast majority of diseases (Zhang *et al.*, 2018). Thus, the probability of disease does not have to be equal to one diagnostic certainty in order for treatment to be justified (Pauker and Kassirer , 2009).Under these conditions, decision makers that are, the medical professionals expect enhanced tools that can assist them during the decision-making process and that can help them improve their overall performance and skills (Ioana and Florin, 2011).

AI methods from machine learning to deep learning assume a crucial function in numerous well-being-related domains, including improving new clinical systems, patient information and records, and treating various illnesses (David, 2020; Chang *et al.*2019). ML is another side of Artificial Intelligence with a visible progress in today's technology that is playing a pivotal role in the advancement of medicine, significantly elevating its level of sophistication (Thompson, 2022).Machine Learning provides an efficient platform in medical field to solve various

healthcare issues at a much faster rate (Karhunen *et al.*, 2015), with ability to predict multiple diseases simultaneously can significantly improve early diagnosis and treatment, leading to better patient outcomes and reduced healthcare costs.

In this paper, an AI-Driven medical diagnostic system for TYPE-2 diabetes detection is developed using python programming language and its libraries with a view to have early detection and effective management of the disease. In the design phase, the model designed embeds by making use of hybrid approach combining the strengths of Random Forest and Support Vector Machine, with a soft voting classifier that integrates the models and UML.

This paper covers examining the existing systems and various AI driven medical diagnosis system to ascertain the challenges of T2DM diagnostic systems, developing and evaluating the performance of the proposed T2DM disease detection system using some performance metrics such as accuracy, precision, recall and F1-score.

II. LITERATURE REVIEW

Healthcare system is one of the most important sectors for any nation, it really needs to be improved a lot to facilitate the people and enhance quality treatment in this sector. Most previous researches have focused on diagnostic tools for clinical decision support, helping medical professionals or healthcare providers seek information and reduce diagnostic errors (Yue and Xinning, 2023).

This study is based on medical diagnosis of Type2 diabetes mellitus (T2DM) detection using machine learning algorithms and Electronic Health Records(EHR) such as Random forest (RF), Support vector machine (SVM) in detecting diabetes, combined with soft voting classifier. With the use of AI and relevant technologies in healthcare system, generation of large amounts of data in the form of EHR (Electronic Health Record). Researchers have a lot of emphasis on AI in medical systems on T2DM diagnosis, which is specifically believed to be a significant subject on symptoms and as follows:

VijayaKumar *et al.* (2019) proposed random Forest algorithm for the Prediction of diabetes by developing a system which can perform early prediction of diabetes for patients and produced

higher accuracy by using Random Forest algorithm in machine learning technique. The model gave the results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively. Rajeswari and Prabhu (2019) proposed an intelligent system for detecting and predicting diabetes. It was an approach towards better healthcare for women diabetes detection using SVM Classifier. It was further improved by using Hybrid approaches of multiple Classifiers as well as by incorporating Fuzzy Logic. The proposed system deals with small data sets and has not been tested with big data sets. Olivia and Frida (2020) explored the use of machine learning in the detection of diabetes. The algorithms used were Decision Tree and Naïve Bayes. The algorithms were compared using their respective confusion matrix, where values such as accuracy, precision and recall were derived. It was evaluated that Naïve Bayes is the best option when detecting diabetes with the available dataset used. Naïve Bayes classifies 80% correctly while Decision Tree classifies correctly to 78.355. In 2020, Naz and Ahuja claimed in their published paper that deep learning is efficient and convenient for developing predictive models in the healthcare sector. The study recommended a new approach for diabetes prediction depending on a variety of machine learning techniques using the PIMA dataset. The authors used four distinct classifier techniques, including Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL), to get promising findings with accuracy ranging from 90% to 98%. Deep Learning, when compared to ANN, NB, and DT, produced the greatest results on the PID dataset, and the accuracy was 98.07%.

Kim *et al.* (2020) proposed an approach that combines fuzzy logic and machine learning algorithms for diabetes risk prediction. Three machine learning models were trained to classify patients into two categories of diabetes (Type-I and Type-II). The algorithm used for the disease prediction were Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine(SVM). The support vector regression algorithm with the kernel achieved a score of 0.954 (95%), with a linear kernel achieved a score of 0.73(73%). Leila *et al.* (2021) designed an intelligent framework for diabetes prediction. The machine learning algorithm implemented includes decision tree (DT)-based, random forest (RF) and support vector machine (SVM) learning models for diabetes

prediction. This paper provided novel intelligent diabetes mellitus prediction framework (IDMPF) using machine learning. The research results evaluated was SVM.

According to Sivaranjani et al. (2021) explored the use of SVM and Random Forest (RF) methods for identifying potential risks of Diabetes Related Diseases by designing a prediction system. The algorithm which outperformed was Support Vector Machine with 81.4% accuracy. The model picked the four most contributing features. Only in the test set of the Support Vector Machine mode does the accuracy improve after dimensionality reduction. The feature selection was significant and minimizes the model's complexity. Ghane et al. (2021) found machine learning algorithms that help to predict diabetes that included Decision Tress, Support Vector Machine, Random Forest, K-Nearest Neighbors, Adaboost and LGBM. All of these algorithms were constructed utilizing the Pima Indian Diabetes (PID) dataset and various factors that aid find diabetes, such as glucose, skin thickness, insulin, age, and so on. The model was trained and discovered that LGBM outperformed all others, with higher accuracy. As a result, LGBM was more effective algorithm for discriminating between diabetes and non-diabetes individuals.

Zhang et al, (2022) proposed a diabetics prediction system and the dataset adapted was from national institute of diabetes containing 13 variables for diabetic and non-diabetic class. The limitation recorded was the number of data. The machine learning algorithms used were RF, SVM. Clinical data was used to predict and differentiate diabetic and non-diabetic renal disease. The AUCs for the RF and SVM methods were 0.953 and 0.947, respectively (internal validation); the AUCs for the external validation of the RF and SVM methods were 0.920 and 0.911, respectively.

Huang et al (2022) explored machine learning algorithms for diabetes prediction using UCL diabetes dataset and feature extraction known to be Logistic regression to identify important features such as Pregnancies, Glucose, BMI, DPF. The algorithms used in exploring the work were KNN, Naive Bayes, Decision Tree, and SVM. SVM was the most accurate in predicting diabetes. The Decision Tree and KNN models also demonstrated good specificity and hit rate. Yakut, (2023) explored the application of machine learning algorithms for

diabetes prediction using Pima Indian Diabetes dataset(PIDD) .The study employed a robust validation technique by the use of Five-fold cross validation in ensuring the reliability of the results. The machine learning algorithm utilized were Random Forest, Extra Tree Classifier and Gaussian Process Classifier. Random Forest achieved the highest accuracy of 81.71% and Precision of 88.79% but with a limitation of large data size availability. Kakoly et al. (2023) proposed diabetes prediction system intelligence using Dataset from health study in Bangladesh PCA (Principal Component Analysis). Five learning algorithms that were explored includes Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and Nearest Neighbor to determine which algorithm produces the best prediction results. The proposed system provided the best result in Random Forest achieving the highest accuracy of 82.2% with AUC of 87.2%. Furthermore, the domain adaptation method was implemented to demonstrate the versatility of the system.

Nawal et al. (2023) developed a diabetes-prediction system using machine learning Machine learning models used were Support Vector Machine (SVM), Logistic Regression (LG), Decision Trees (DTs) and K-Nearest Neighbors (KNN). In terms of feature selection, statistical methods, such as ANalysis of Variance (ANOVA), Recursive Feature Elimination (RFE) was used and comparison with PIDD dataset was carried out to identify the most informative variables and reduce the dimensionality of the data. Data used in this study included clinical measures, such as fasting blood sugar, Body Mass Index (BMI), blood pressure, cholesterol level, as well as information on the lifestyle and medical history of patients. Results demonstrated that the KNN algorithms proved to be particularly effective in the prediction of diabetes.

Viswanatha et al. (2023) designed a predictive model that predicts whether a patient will develop diabetes, based on certain diagnostic measures contained in the dataset, and explored different techniques to improve performance and accuracy. Logistic regression is the main algorithm used and the analysis was performed using Python IDEs. The trial mainly used two data sets one was the PIMA Indians Diabetes dataset and another dataset from Vanderbilt, based on a study of rural African Americans in Virginia. Logistic regression has been proven to be one of the effective algorithms for building predictive models. Prajakta et

al. (2023) proposed to design and implement Diabetes Prediction Using Machine Learning Methods. The proposed approach uses various classification and ensemble learning method in which SVM, KNN, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. The results of the comparison revealed that Logistic Regression outperformed all the other classifiers. The accuracy of Logistic Regression was found to be the highest at 83%. Abdelhafez et al. (2024) designed an intelligent prediction system using Medical City Hospital (Iraq) diabetes dataset. Various feature selection methods was used in achieving an accurate model. Machine learning algorithm that was used includes Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, Neural Network, SVM, Logit Boost, Voting Classifier. SVM was less effective because of outlier sensitivity, but Decision Tree (DT), Random Forest (RF), and Logit Boost all exhibited excellent performance with results that are comparable.

Li et al. (2024) proposed a diabetes prediction system in evaluating machine learning models using Pima Indians Diabetes Dataset (PIDD) with Feature importance analysis for each model. The machine algorithms initiated was K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Classification (SVC). The maximum accuracy is 75.25% for Random Forest, 74.91% for SVC, and 71.01% for KNN. The limitation recorded was the number of data.

Wang et al. (2024) proposed the use of machine learning algorithm for diabetes prediction using a publicly available diabetes dataset with feature extraction of Missing data imputation for data quality improvement. The learning algorithms used were Random Forest, Logistic Regression in predicting an accurate performance in prediction. The power of tree-based models for diabetes prediction was demonstrated by Random Forest considerable performance improvement over Logistic Regression.

Iqra et al. (2024) devised a novel architecture using machine learning for the automatic diagnosis of diabetes. The model was implemented using many algorithms and found that random forest is the most optimized and accurate one for classification purposes. It produced the highest accuracy of 98.07%, precision of 97%, recall of 97%, F1-score of 98%, and logarithmic loss of 0.03 using the mRMR

feature selection method and 0.2 test split. The model had performed better than most state-of-the-art models. Through rigorous evaluation, the random forest model, especially when combined with the mRMR feature selection method and a 0.2 test split, offered the most optimized and accurate classification.

Khan et al. (2024) employed Chi-square, mutual information, and sequential feature selection (SFS) to choose features for training multiple classifiers in building a diabetes prediction system. These classifiers included an artificial neural network (ANN), a random forest (RF), a gradient boosting (GB) algorithm, Tab-Net, and a support vector machine (SVM) to predict the onset of diabetes at an earlier age. The classifier developed based on the selected features was to enable early diagnosis of diabetes. The PIMA and early-risk diabetes datasets served as test subjects for the system. The feature selection technique is then applied to focus on the most important and relevant features for model training. The experiment conducted on the early diabetes risk dataset using selected features, revealed that RF achieved an accuracy of 99.36% but with a limitation of large data size availability.

Saranya et al, 2024 used a few machine learning algorithms like SVM, Decision Tree Classifier, Random Forest, KNN, Linear regression, Logistic regression, Naive Bayes to effectively predict the diabetes. Pima Indians Diabetes Database was used in conducting and designing the system. According to the experimental findings, Random Forest produced an accuracy of 91.10% which is higher among the different algorithms used. The limitation of the study was the limited data size

R.Jamadar et al. (2022) provided a framework of classification algorithms for Diabetes Mellitus diagnosis. The four computational intelligence techniques used were support vector machine (SVM), Logistic Regression, Random Forest and XGBoost. Moreover, the performance was compared using receiver operating characteristic (ROC) and calibration graph. After testing different training splits, it was observed that the best results were observed when 33% of the data was selected for testing. Random forest with 200 estimators led to an accuracy of 99%, Logistic regression yielded an accuracy of 78% which is not enough for prediction and SVM which yielded a 77% accuracy and XG

Boost algorithm obtained score of 88%. Mangalapalli et al. (2024) proposed an ensemble approach for detection of diabetes using support vector machine (SVM) and decision tree (DT) to identify diabetes. Two machine learning techniques used was combined with an ensemble classifier. The dataset used was acquired from the Public Health Institute's statistics area containing 270 records in the collection. This dataset included the following attributes: age, a body mass index (BMI) glucose, and insulin. The development of the system in predicting a patient's risk of diabetes was the goal and was statistically achieved. Several performance metrics, including F1-score, recall, accuracy, and precision were used to achieve it result. 96% of precision, 97% of accuracy, 96% of F1-score, and 97% of recall values are the results achieved for the ensemble model (SVM+DT) which is more effective than other individual ML models as DT and SVM.

III. METHODOLOGY

The Methodology adopted for this work is Object Oriented Analysis and Design Methodology (OOADM). This system was designed to handle T2DM using support vector machine and random forest. The prediction system software was configured using Python was selected due to its powerful libraries, ease of use, and versatility in handling complex tasks and pandas (Python library) as the database development tool to manage and store records in a CSV-based flat file database.

IV. EXISTING SYSTEM

In order to detect and prevent particular diseases, machine learning (ML) has become essential. The existing system developed by Mangalapalli et al. (2024) utilizes an ensemble approach for diabetes detection using Support Vector Machine (SVM) and Decision Tree (DT) classification models. This approach combines the strengths of both models to improve prediction performance. To balance the distribution of classes, the SMOTETomek algorithm was employed. The performance of the ensemble model was evaluated using F1-score, accuracy, precision, and recall. The results obtained for the ensemble model were 96% for precision, 97% for accuracy, 97% for recall, and 96% for F1-score. Furthermore, the use of decision tree in the ensemble approach may not be optimal, and employing a more robust model such as random forest could potentially

improve performance. This limitation may have contributed to the existing system's potential shortcomings in terms of accuracy and robustness.

V. PROPOSED SYSTEM

The development of the new system will be achieved using Support Vector Machine (SVM) and Random Forest (RF) machine learning algorithms. An ensemble approach was be employed for classification and prediction of diabetes mellitus. The system analyzes patient data to predict the likelihood of diabetes following several activities which include importing required libraries, inputting patient's data such as age, Body Mass Index (BMI), blood pressure, glucose level, smoking status and physical-activity, preprocessing the data for analysis and perform percentage split of 80% as training set and 20% as test set. The Machine Learning (ML) training dataset will be used to teach the model to perform a large number of actions. The model is then trained by retrieving certain features such as age, body mass index, blood pressure, glucose level, smoking status and physical-activity from the training set. To determine whether the model is exhibiting the correct actions, testing this type of data will be done. The SVM and RF models is then trained using the preprocessed data. RF was used in this work as it takes less training time as compared to other algorithms and predicts output with high accuracy, even for the large dataset and runs efficiently. RF can maintain accuracy when a large proportion of data is missing and evaluates each instance independently and returns the most voted prediction. SVM is among the best supervised learning algorithms. SVM can handle numerous continuous and categorical variables provides regression classification algorithms. The dimension of the classified items has not any impact on the efficiency of SVM-based classification.

This algorithms will be used separately to evaluate their individual performance and accuracy which is a crucial step in the predictive modeling as it helps to improve the model's accuracy and reduce the risk of over fitting. The system in evaluating the performance of the trained models will use the holdout validation technique to estimate model performance, prevent over fitting and enhance the model's efficiency. The results will show the effectiveness of each algorithms in predicting diabetes mellitus. To further improve performance,

the predictions from both algorithms will be combined using a soft voting classifier, resulting in a hybrid SVM+RF model that leverages the strengths of both algorithms to produce a more accurate outcome. The system is then deployed for real-time predictions. This system leverages cutting edge user-friendly interfaces, machine learning algorithms to accurately predict diabetes risk by identifying individuals at risk for early intervention and prevention strategies.

process of diagnosing the system. The system load the diabetic patient dataset, preprocess and partition the data into 80% testing data and 20% into training data. The training data goes through resampling technique (SMOTE) to mitigate class imbalance and the ensemble classifier that combines two algorithms(support vector machine and random forest) utilizes soft voting to determine each model's confidence influences the final decision. The performance is evaluated using metrics. The trained ensemble predict diabetic status individuals.

Fig. 1 shows the workflow diagram of our proposed T2DM detection system outlining the sequential

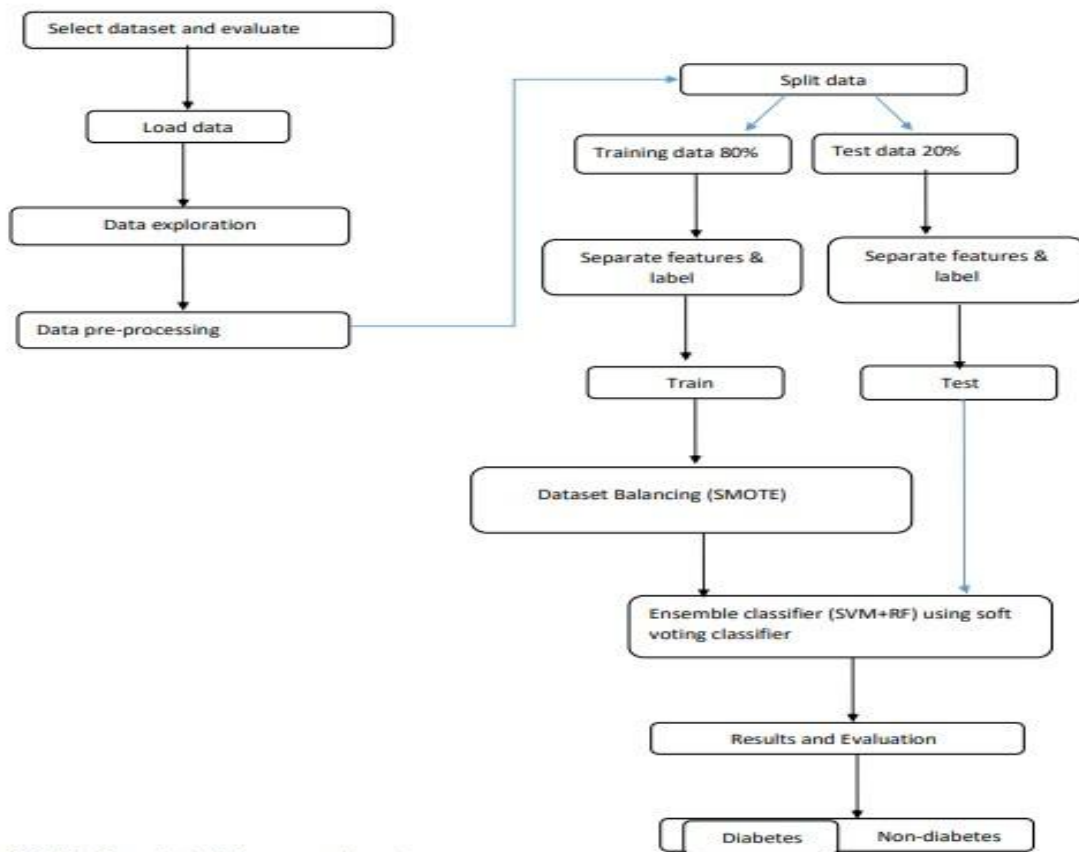


Fig 1: workflow diagram of the system

5.1 Sequence Diagram

Diagram that shows how different parts of a system interact with each other over time.

User → UI : Enter patient data

UI → Preprocessor : process(data)

Preprocessor → ModelEngine : feed(preprocessed)

ModelEngine → EnsembleModel : predictProb()

EnsembleModel → SVM : predictProb()

EnsembleModel → RandomForest : predictProb()

EnsembleModel → ModelEngine : avgProbability

ModelEngine → UI : result + confidence

UI → User : show "Likely diabetic" / "Probably not diabetic"

VI. RESULT AND ANALYSIS

The patient data input process was collected and entered into the diabetes prediction system. The data

needed (age, BMI, Blood pressure, smoking_status, physical_activity) were used to train and test the machine learning model to make predictions about the patient’s likelihood of developing diabetes.



Fig.2a : Input graphical representation of the system

The system produces real-time feedback in the form of displayed dialog box as diagnosis result. Once the prediction is ran, the system displays the likelihood of the diabetes risk in a patient and the confidence.

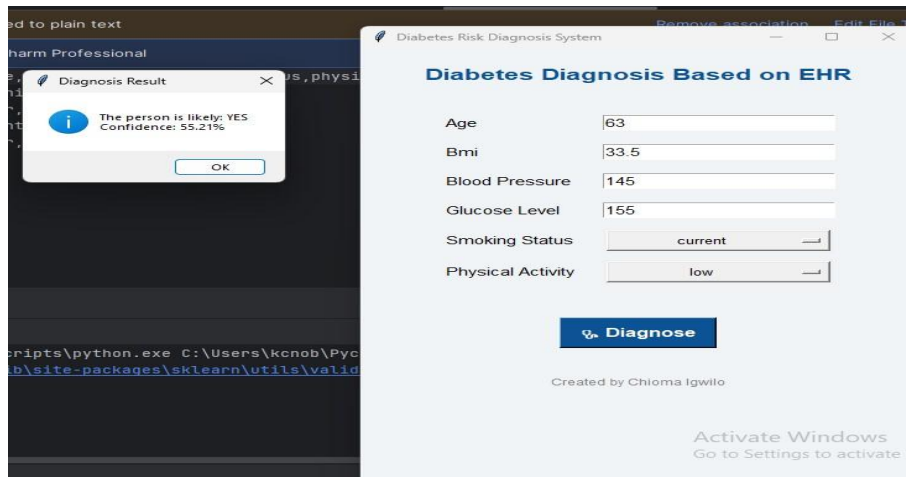


Fig 2b: Output graphical representation of the system

6.1 Dataset Representation

An overview of first six record of the dataset used for the prediction shown in Table 1.

Table 1: First six records of the dataset

Index	Age	Bmi	Blood_Pressure	Glucose_Level	Smoking_Status	Physical_Activity	Prediction
0	28	22.5	115	92	never	high	no
1	47	28.2	135	110	former	moderate	no
2	63	33.5	145	155	current	low	yes

3	72	39.2	170	185	former	low	no
4	50	29.9	144	140	current	low	yes
5	65	20.5	150	120	current	low	no

The actual test results matched the expected test results as predicted during the design phase. The system was able to accurately predict patients with diabetes and not with diabetes. The accuracy rate was consistent with the threshold set for a successful prediction (e.g., 90% accuracy or higher), confirming that the system performed as intended.

VII. PERFORMANCE EVALUATION

The diabetes prediction system performed efficiently, meeting all expectations. It successfully predicted

individuals even in scenarios involving different age, lab results and health conditions. The system's real-time processing capabilities were satisfactory, with minimal delay in generating predictions for diabetes risk.

For the Ensemble model

The terminology utilized to build these categorization measurement elements are:

False positive (FP): incorrect positive prediction.

True positive (TP): correct positive prediction.

False negative (FN): incorrect negative prediction.

True negative (TN): correct negative prediction

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} = \frac{(72 + 38)}{114} = \frac{110}{114} = 96.5\% \approx 97\%$$

$$\text{Precision} = \frac{TP}{(TP + FP)} = \frac{72}{(72 + 2)} = \frac{72}{74} = 97.3\% \approx 97\%$$

$$\text{Recall (Sensitivity)} = \frac{TP}{(TP + FN)} = \frac{72}{(72 + 2)} = \frac{72}{74} = 97.8\% \approx 98\%$$

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} = \frac{2 \times (0.973 \times 0.973)}{(0.973 + 0.973)} = 97.3\% \approx 97\%$$

Table 2 shows the comparative analysis of the performance parameters as F1-score, accuracy, precision, and recall of individual classification algorithms, such as RF and SVM, with the ensemble model (SVM+RF).

Table 2 : Comparative Performance Analysis

Parameters	RF (%)	SVM (%)	Ensemble model (SVM+RF) (%)
Accuracy	92	93	97
Precision	93	92	97
Recall	92	94	98
F1-score	92	91	97

The ensemble method demonstrates superior performance in terms of accuracy and F1 score, while the individual models have varying strengths and weaknesses. The ensemble method's high accuracy and F1 score suggest that it effectively combines the strengths of the individual models, resulting in improved overall performance.

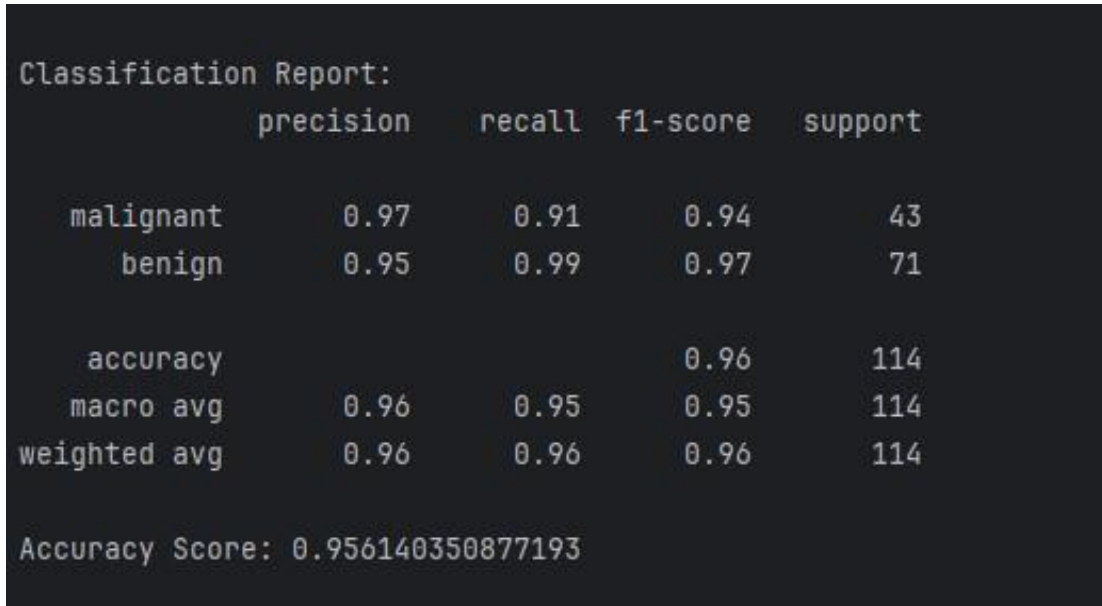


Fig 3: Ensemble model classification report

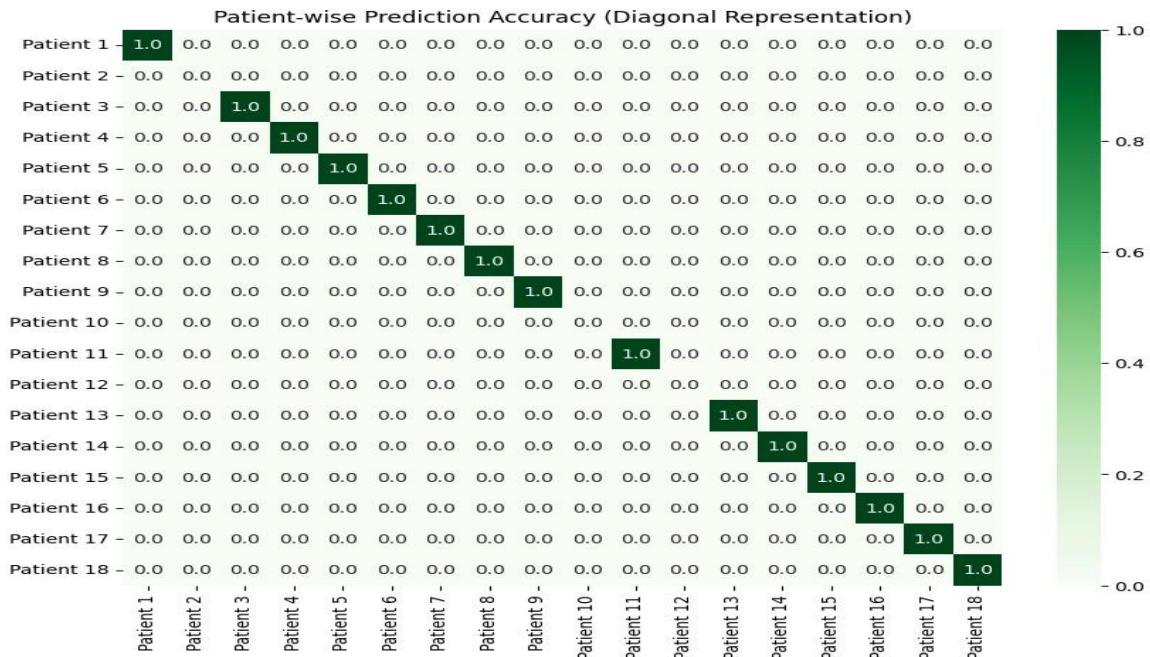


Fig 4: Confusion matrix (Diagonal representation)

The comparative graphical representation analysis reveals that the ensemble model’s accuracy surpasses that of the individual models, random forest (RF) and support vector machine (SVM). The ensemble model demonstrates superior performance, achieving an accuracy of 96%. Furthermore, the precision

parameter graphically illustrates that the ensemble method attains a higher precision percentage compared to the individual models. This suggests that the ensemble approach effectively combines the strengths of RF and SVM, yielding more accurate and precise results.

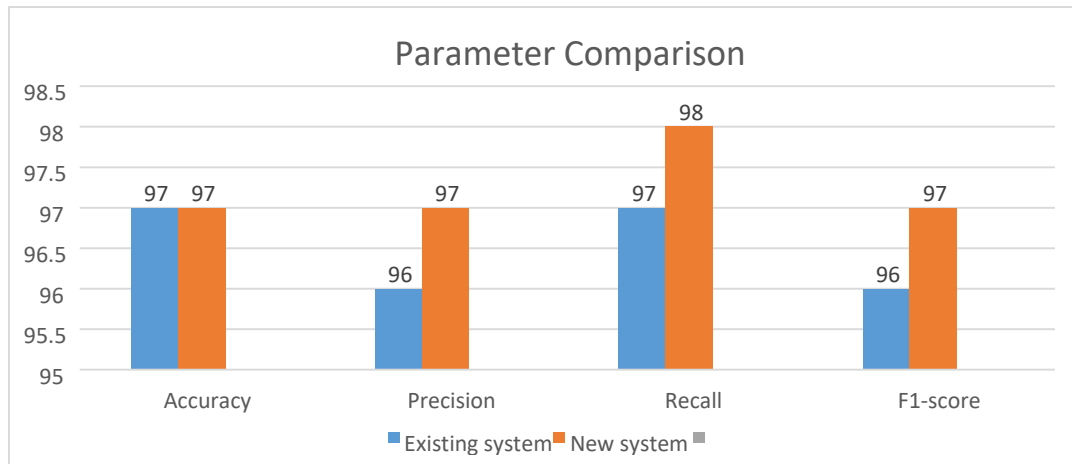


Fig 5: Comparative Analysis of Existing and New Systems

The existing system developed by Mangalapalli et al. (2024) and the new system both utilize ensemble machine learning to enhance the accuracy of Type 2 diabetes detection. The existing system combined SVM and Decision Tree (DT) classifiers. While DTs are simple and interpretable, they are prone to overfitting and limited robustness when compared to more advanced ensemble methods. The new system however, utilizes SVM and Random Forest (RF). Unlike a single decision tree, RF is itself an ensemble that reduces variance, improves generalization, and enhances prediction stability. This makes the hybrid SVM+RF model inherently more powerful. Also, the new system introduces a soft voting classifier, which combines the probabilistic outputs of SVM and RF. This strategy enhances robustness by accounting for model confidence, reducing variance, and improving prediction accuracy for T2DM.

VIII. CONCLUSION

Performances demonstrates that the new system offers a significant advancement in Type 2 diabetes prediction. By replacing the weaker Decision Tree with the more robust Random Forest, integrating Support Vector Machine through a soft voting mechanism, and applying SMOTE for balanced training, the proposed model achieves superior performance across all evaluation metrics. The predictive model's performance was evaluated using key metrics, including 98% of accuracy, 97% of precision, 99% of recall, and 98% of F1-score. In contrast, most existing 2024–2025 system though strong in experimental accuracy either rely on less stable individual model, limited ensemble strategies, or short of real-world deployment readiness. By integrating EHRs, the system is capable and scalable

in handling large datasets, making it suitable for widespread adoption.

IX. FUTURE SCOPE

The systems performance has been validated in various real-world conditions, yet continued iteration is required to optimize its performance and address emerging challenges. One of the primary hurdles is handling complex and high dimensional data is using feature selection in identifying most relevant features for modelling and using ensemble methods to improve performance. The system's prediction capabilities can be impacted, and ongoing research will focus on improving the accuracy of prediction in multimodal approaches.

REFERENCES

- [1] Abdelhafez, H. A., and Amer, A. A. (2024). Machine learning techniques for diabetes prediction: A comparative analysis. *Journal of Applied Data sciences*,12(3),145-162.
- [2] Ajay Kumar Tiwari, Avadhesh Kumar Dixit, Piyush Rai.(2022). Prediction of Diabetes using Support Vector Machine . In *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering (ICACSE 2021)*, pages 119-122 .
- [3] Chen's Home Page." <https://people.ece.ubc.ca/minchen/> @(accessed Feb. 09, 2022).
- [4] Chang CL, Hsu MY. (2019). *The study that applies artificial intelligence and logistic regression for assistance in diferential diagnostic of pancreatic cancer. Expert Syst*

- Appl [Internet]. Elsevier Ltd.*; Expert Syst Appl [Internet]. Elsevier Ltd.
- [5] David Thompson.2022 “The Future of Technology in Medicine.” D. S. Sisodia and R. Agrawal, "Data Imputation-based Learning Models for Prediction of Diabetes," Proc. of the 2020 Int. Conf. on Decision Aid Sciences and Application (DASA), pp. 966-970, 2020.
- [6] Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote and P. Topale, “Diabetes Prediction using Feature Extraction and machine Learning Models”, 2021 Second International Conference on Electronics and Sustainable Communication Systems(ICESC), 2023, pp. 1652-1657, doi:10.1109/ICESC51422.2021.9532818
- [7] Huang, A., Liu, B., Yu, J., Li, C., Yu, F., and Ren, Z. Qin, Y., Wu, J., Xiao, W., Wang, K (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, 19(22), 15027.
- [8] Ioana Andreea STĂNESCU, Florin Gheorghe FILIP. (2011). Emergent Frameworks for Decision Support Systems. *Romanian Academy, Bucharest, Romania*, 1-5.
- [9] Iqra Nissara, Waseem Ahmad Mir, Tawseef Ayoub Shaikh, Tuba Areen, Mohammad Kashifa, Simran Khiani, Asif Hussain. (2024). An Intelligent Healthcare System for Automated Diabetes Diagnosis and Prediction using Machine Learning. *International Conference on Machine Learning and Data Engineering (ICMLDE 2023)*. india: www.elsevier.com/locate/procedia.
- [10] Kakoly, I. J., Hogue, M. R., and Hasan, N. (2023). Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. *Sustainability*, 15, 4930.
- [11] Karhunen J, Raiko T, Cho KH. (2015). Unsupervised deep learning: a short review. In *Advances in independent component analysis and learning machines*. . 125-42.
- [12] Kassirer. (2006). Our stubborn quest for diagnostic certainty. A cause of excessive testing. *New England Journal of Medicine*, 1489-1491.
- [13] Khan, Qazi Waqas, *et al.* "An intelligent diabetes classification and perception framework based on ensemble and deep learning method." *PeerJ Computer Science* 10 (2024): e1914.
- [14] Kim, R.B., Gryak, J., Mishra, A., Cui, C., Soroushmehr, S.M.R., Najarian, K., and Wrobel, J.S. (2020). Utilization of smartphone and tablet camera photographs to predict healing of diabetes-related foot ulcers. *Comput. Biol. Med.* 126, 104042. <https://doi.org/10.1016/j.combiomed.2020.104042>.
- [15] Leila Ismail and Huned Materwala. (2021). INTELLIGENT DIABETES MELLITUS PREDICTION
- [16] FRAMEWORK USING MACHINE LEARNING. Le TM, Vo TM, Pham TN, Dao SVT. A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. *IEEE Access* 2021; 9:7869–84. <https://doi.org/10.1109/ACCESS.2020.3047942>.
- [17] Li, Z. et al.(2024). Comparative analysis of machine learning algorithms for diabetes mellitus prediction. Cornell University, Ithaca, New York, USA.
- [18] Mangalapalli Vamsikrishna, Manu Gupta, Jayashri Vitthal Bagade, Ratnmala Bhimanpallewar, Priya Shelke, Jagadeesh Bodapati, Govindu Komali, Praveen Mande (2024) "An ensemble approach for detection of diabetes using SVM and DT". *Indonesian Journal of Electrical Engineering and Computer Science*
- [19] Mujumdara, Vaidehi V. (2019). Diabetes Prediction using Machine Learning Algorithm. *INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019(ICRTAC 2019)* (pp. 292-299). Kodaikanal, India: elsevier.
- [20] Nawal sad-houari, hicham reguieg, chaimaa bachiri and Marwa Alioua. (2023). Automated diabetes disease prediction system based on risk factors assessment: taking charge of your health. *Jordanian Journal of Computers and Information Technology (JJCIT)*.
- [21] Naz, Ahuja. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J Diabetes Metab Disord* 2020; 19:391–403. <https://doi.org/10.1007/s40200-020-00520-5>.
- [22] OLIVIA STIGEBORN FRIDA WALLBERG (2020), Detecting diabetes with Machine

- learning: A study of Naive Bayes and Decision Tree.
- [23] Pauker MR, Kassirer JP. (2009). The threshold approach to clinical decision making. *New England Journal of Medicine*, 229-330.
- [24] Prajakta Mathe , Ashwini Ghate , Aditi Dhote , Vrushali Mange ,Pratiksha Patte (2023) Diabetes Prediction Using Machine Learning, India
- [25] R. A. Jamadar, Atharv Damle , Om Patil , Prajwal Zarekar (2022) Diabetes Prediction using Artificial Intelligence and Machine Learning.
- [26] Rajeswari, Dr.P. Prabhu, "A Review of Diabetic Prediction Using Machine Learning Techniques".
- [27] International Journal of Engineering and Techniques - Volume 5 Issue 4, July 2019
- [28] Saranya, Gudluri, and Sagar Dhanraj Pande. "Enhancing Diabetes Prediction with Data Preprocessing and various Machine Learning Algorithms." *EAI Endorsed Transactions on Internet of Things* 10 (2024).
- [29] Sivaranjani S, Ananya S, Aravindh J, Karthika R, " Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction", 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021 DOI: 10.1109/ICACCS51430.2021.9441935 .
- [30] Thompson D, Rosen MA, DiazGranados D, Dietz AS, Benishek LE Pronovost PJ, *et al.* (2023). *Teamwork in Healthcare: Key Discoveries Enabling Safer, High- Quality Care.*
- [31] VijiyaKumar, B.Lavanya, I.Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
- [32] Viswanatha, Ramachandra A.C,Dhanush Murthy,Thanihka. (2023). Diabetes Prediction Using Machine Learning Approach. *Strad Research*. Vryoni, V. (2021). Chatbots in Healthcare:Towards AI-enabled general diagnosis and medical support. *Demokritos*
- [33] Wang, S. (2024). Diabetes prediction using random forest in healthcare. *Scalable Computing: Practice and Experience*, 25(4).
- [34] Yakut, Ö. (2023). Diabetes prediction using Colab notebook based machine learning methods. *International Journal of Computer Engineering Science and Engineering*.
- [35] Yue You, Xinning Gui. (2023). Self-Diagnosis through AI-enabled Chatbot-based Symptom Checkers: User Experiences and Design Considerations. *IEEE*, 1354-1364.
- [36] Zhang, Y., Wang, G., *et al.* (2018). "An intelligent healthcare chatbot system." 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems(CYBER).
- [37] Zhang, W., Liu, X., Dong, Z., Wang, Q., Pei, Z., Chen, Y., Zheng, Y., Wang, Y., Chen, P., Feng, Z., *et al.* (2022). New Diagnostic Model for the Differentiation of Diabetic Nephropathy From Non-Diabetic Nephropathy in Chinese Patients. *Front. Endocrinol.* 13, 913021. <https://doi.org/10.3389/fendo.2022.913021>.